ERIC Document Resume	See Instruction Sheet	DATE OF RESUME 6/26/65	DATE RECEIVED

A. PUBLICATION DATA

1. ED NUMBER	2. ORIGINATOR CONTROL NO(S).	3. IDENTIFICATION NO.
ED001122	N.A.	CR 2234

4. AUTHOR(S): Stodola, Quentin C., Eustice, D. Edward, and Kolstoe, Ralph

5. TITLE OF DOCUMENT ABSTRACTED:

Frequent Classroom Testing as a Learning Aid Using Data Processing

6. PUBLICATION TITLE: N.A.

7. EDITOR(S): N.A.

8. PLACE OF ORIGIN
ORG.
ST. North Dakota State University
CITY, STATE
TELE. Fargo, N.D.
INST/LOC CODE

9. PUBLISHER OR SOURCE Cooperative Research Program, Office of Edu., Dept. of HEW Washington, D. C.

10. VOL.	11. PAGES	12. MO./YR.	13. TOTAL PP.
N.A.	N.A.	Nov. 1964	64

DO NOT WRITE IN SPACE BELOW

MICROFILM COPY AVAILABLE YES ☒ NO ☐

COPYRIGHT GRANTED YES ☐ NO ☒

COPYRIGHT RELEASE FILED YES ☐ NO ☐

TYPE OF REPORT
PROPOSAL ☐
IN-PROGRESS ☐
FINAL ☒
FOLLOW-UP ☐

OTHER ERIC REFERENCES:

ERIC SATELLITE CODE

B. INDEXING DATA

14. SUBJECT AREA

15. TYPE OF ENDEAVOR report

16. INDEXING TERMS (6-12)

testing procedures	test evaluation		
data processing	achievement tests	ability tests	
teaching methods	interest inventories	university level	

17. ABSTRACT (MAX. 250 WORDS)

The use of practice tests as a learning aid were evaluated. Attempts were made to demonstrate how data processing equipment could be used effectively to reproduce tests and analyze test results for improvement of instruction.

Initially, a class of students in tests and measurements at North Dakota State University (the reference group) was administered five unit tests and a final examination. The teaching procedures followed the usual lecture-discussion method. A second class in tests and measurements (the experimental group) was taught by a different method but administered the same unit tests and final examination as the reference group. Lectures were compressed to permit time for a practice test followed by discussion to precede each unit test. The practice tests were parallel in content and difficulty with the unit tests.

The data indicated higher mean scores for the group which was administered the practice tests. Forty-seven out of 48 different comparisons favored the experimental group when separate consideration was given to various subgroups. A large majority of the students favored the use of practice tests followed by discussion of results. The data processing equipment provided an effective means for reproducing test booklets and served as an aid for planning instruction to cover points not fully understood by the class. Overall, use of practice tests and data processing equipment is believed to have contributed to effective learning in this experimental study.

ERIC No.

18. SOURCE OF ABSTRACT

C. FUNDING DATA

19. SUPPORTING ORGANIZATION	20. DIVISION OR DEPT.	21. CONTRACT NO. / PROJECT NO.
	22. TOTAL COST OF PROJECT	
	23. DURATION OR PROJECT	U.S. Department of Health, Education, and Welfare Office of Education

ERIC Form 1

REPRINTED BY PERMISSION
U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN PRINTED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Cooperative Research Project 2234

Frequent Classroom Testing As A Learning Aid Using Data Processing

Prepared by Quentin C. Stodola, Northern Michigan University (formerly of North Dakota State University); D. Edward Eustice, North Dakota State University; Ralph H. Kolstoe, University of North Dakota

The research reported herein was supported through the Cooperative Research Program of the Office of Education, U.S. Department of Health, Education and Welfare.

Acknowledgments

Appreciation is expressed to Donald E. Peterson, Director of the Data Processing Laboratory, for computational services, and Beth Rochefort for assistance in editorial planning, both on the staff of North Dakota State University.

TABLE OF CONTENTS

	Page
USING TESTS FOR BETTER LEARNING	5
The Problem	5
Related Research	5
PROCEDURES	9
The Sample	9
The General Plan	10
Constructing the Tests	11
Teaching Methods	13
RESULTS	15
Comparison of Reference and Experimental Classes Considered as a Whole	15
Percent Correct on Individual Items by Groups	18
Comparative Difficulty of Practice and Unit Tests	19
Intercorrelations	19
Scores Obtained by Students with High and Low Ability as Measured by SCAT	19
Scores Obtained by Men and Women Considered Separately	19
Relative Success of Men and Women Under the Two Teaching Conditions	24
Student Attitude Toward Practice Tests	24
DISCUSSION OF RESULTS	27
Results for Total Group	27
Results for Subgroups	28
Difference in Percent Correct for Reference and Experimental Groups on Individual Items	28
Difficulty of the Unit Tests	29
Attitude of Students Toward Practice Testing	30
USE OF THE DATA PROCESSING EQUIPMENT	31
Reproducing the Tests	31
Test Scoring and Item Analysis	32
SUMMARY	33
APPENDIX	35

LIST OF TABLES AND APPENDICES

	Page
TABLE 1	9
Characteristics of Groups According to Class, College, and Sex	
TABLE 2	10
Comparison of the Means and Standard Deviations of Experimental and Reference Groups on SCAT	
TABLE 3	11
Categories Within Classification System	
TABLE 4	15
Means and Variances of Test Scores of the Reference, Experimental, and Combined Groups	

LIST OF TABLES AND APPENDICES—(Continued)

TABLE 5 16
Analyses of Variance on Unit Tests and Final Examination for Total Reference and Experimental Groups

TABLE 6 17
Summary of Analyses of Variance and Covariance of Test Scores for the Total Group

TABLE 7 18
Number of Items Per Test on Which Reference and Experimental Differ at the .05 Level in Percent of Correct Responses

TABLE 8 20
Summary of Analyses of Variance and Covariance Results for Upper One-Fourth on SCAT

TABLE 9 21
Summary of Analyses of Variance and Covariance Results for Lower One-Fourth on SCAT

TABLE 10 22
Summary of Analyses of Variance and Covariance for Men

TABLE 11 23
Summary of Analyses of Variance and Covariance for Women

TABLE 12 25
Summary of Analyses of Variance and Covariance for Lower Half of Men on SCAT Scores

TABLE 13 26
Teaching Method Preferred by Students at Beginning and End of Course

APPENDIX A 35
Major Subject Classifications and Sub-Classifications of the Item Coding System

APPENDIX B 42
Items in Practice, Unit, and Final Examination Classified According to Subjects Covered

APPENDIX C 44
Items in Practice, Unit and Final Examination Classified According to Skill Required for Response

APPENDIX D 45
Practice, Unit, and Final Examinations According to Difficulty and Biserial r

APPENDIX E 46
Individual Tutoring Given to Reference and Experimental Groups by Sex

APPENDIX F 46
Differences Between Practice and Unit Tests

APPENDIX G 47
Intercorrelation Among Test Scores

APPENDIX H 49
Summary of Analyses of Variance and Covariance Results for Reference and Experimental Groups Separated into Men and Women

APPENDIX I 51
Samples of Items in Which Either Experimental or Reference Group Scores Significantly Higher Than the Other

USING TESTS FOR BETTER LEARNING

The Problem

The major purpose of teaching should be to help students learn. For teaching to be effective there is a need to assess student learning which may be accomplished through testing. Such testing requires great care in the preparation of test items and detailed analysis of student responses.

Unfortunately, classroom tests have too often been used merely to obtain an over-all average which results in assigned grades. For tests to be of most value in the learning process, detailed analysis of test results is necessary so that difficult facts or concepts may be re-emphasized.

Such analysis has in the past not been practical. The development of modern data processing equipment appears to hold the answer in providing detailed information on student learning with a reasonable expenditure of teacher time.

The present research has two major purposes:

(1) to explore the use of generally available data processing equipment for preparation and analysis of classroom tests.

(2) to evaluate the effectiveness of teaching procedures which utilize the detailed information provided by data processing equipment.

Related Research

Much advantage has been claimed for the frequent use of carefully planned classroom tests.

Tyler[1] states that pupils

> ". . . should have continuing experience with appropriate tests to gain confidence in their ability to learn and to perform and to utilize the re-enforcement potential of the tests as important dynamic factors in learning."

[1]*Tyler, Ralph W. "What Testing Does to Teachers and Students," 1959 INVITATIONAL CONFERENCE ON TESTING PROBLEMS, p. 16, October 31, 1959.*

The Committee on Measurement and Evaluation of the American Council on Education[1] makes the following comment:

> "When students are given an opportunity to test their learning by responding to test questions or other evaluative devices, when they learn whether or not their responses are correct, and when they have an opportunity to correct misinformation, there is greater advance in learning then when they lack these opportunities."

However, the advantages claimed for frequent testing seem to depend more on hypothesis than on sound empirical evidence. A review of the literature does not provide a clear-cut answer to the question of whether or not frequent testing improves learning.

In the 1930's, a number of studies were carried on attempting to measure the effects of frequent testing. Ross and Stanley[2] list sixteen studies and Noll[3] six others in addition. The large majority of the studies involved college students, particularly those in classes in general and educational psychology. However, a few were conducted at the high school level. There is no mention of studies carried on at the elementary or junior high school level.

With the exception that the criterion scores in most of these studies were final examinations or end of unit tests, the studies were quite diverse in their procedures. In only two of the studies apparently was there a feed-back of correct answers to the students followed by discussion based on test items.

Generally speaking, the results of the studies listed by Ross and Stanley and by Noll were as expected. Those who were given frequent tests tended to do better on criterion tests. However, in only about a third of the studies were results significant at or beyond the .05 level. It also should be pointed out that in two of the studies, the reference groups, which were not given the frequent tests, did better on the criterion tests than did the experimental.

Four of the studies indicated that the less able students were helped more by frequent testing than were the others. In this connection, Noll says regarding frequent testing, "If tests are used as a frequent basis for specific remedial instruction, results seem beneficial."[4]

Of some interest is student reaction to frequent testing. Concerning this point, Ross and Henry state, "Upon one point, however, the evidence is most conclusive. Jones, Turney, Keyes, and Noll all

---

[1]*Committee on Measurement and Evaluation of the American Council on Education COLLEGE TESTING, p. 43, American Council on Education, Washington, D.C., 1959.*

[2]*Ross, C. C. and Julian C. Stanley, MEASUREMENT IN TODAY'S SCHOOLS, Prentice-Hall, Inc., New York, 1955*

[3]*Noll, Victor H., "The Effect of Written Tests upon Achievement in College Classes: An Experiment and a Summary of Evidence," JOURNAL OF EDUCATIONAL RESEARCH, 32: 345-358, January, 1939.*

[4]*Noll, ibid., p. 348.*

thought at least 70 percent of the students favored the frequent tests to the infrequent and thought that they had learned more because of them. Contrary to a popular belief, many short tests add to the student's enjoyment of a course; if not to his education."[1]

More recently there have been a few other studies relating to the advantages of frequent testing. A study by Standlee and Popham[2] indicates that the greatest gain in learning comes at the beginning of a course and declines by the end of the course. Selakovich[3] found that there was no improvement in learning from frequent testing in a college course in American government. In studies reported by Patterson, Denny and Feldhusen[4] at Purdue University, conducted with psychology students, no improvements were found in frequent testing, although the students themselves seemed to favor this method as contrasted to testing occasionally.

In all of the research described thus far, there does not appear to have been any special effort to use frequent testing as a method of identifying troublesome points which need to be clarified and then as the basis for retesting to help assure that the students' knowledge is complete. There is no mention specifically of item analysis results used as a basis for remedial teaching.

In the past few years, considerable attention has been given to the effects of coaching and practice testing as they may affect scores on tests of general ability. And lately there has been much interest in the effectiveness of programmed instruction, some types of which employ a multiple-choice questions approach.[5] But research efforts of this type have been peripheral to the question of whether or not frequent classroom testing improves learning.

Some attempts have been made to establish stockpiles of test questions for use by classroom teachers. In connection with the publication of various textbooks, test items have been prepared for use in instructor's manuals. However, in the opinion of the investigators, these test items have been generally limited as to both quantity and quality. What appears to be a fine group of test items is the QUESTIONS AND PROBLEMS IN SCIENCE: TEST ITEMS No. 1 developed by Dressel and Nelson.[6] In the specific field of tests and

---

[1]*Ross, C. C. and Lyle K. Henry, "The Relation between Frequency of Testing and Progress in Learning Psychology," JOURNAL OF EDUCATIONAL PSYCHOLOGY, 30: 604-611, November, 1939.*

[2]*Standlee, Lloyd S., and W. James Popham, "Quizzes Contribution to Learning," THE JOURNAL OF EDUCATIONAL PSYCHOLOGY 51: 322-325, 1960.*

[3]*Selakovich, Dan, "An Experiment Attempting to Determine the Effectiveness of Frequent Testing as an Aid to Learning in Beginning College Courses in American Government," THE JOURNAL OF EDUCATIONAL RESEARCH 55: 178-180, January, 1962.*

[4]*Paper given at Annual Meeting of National Council on Measurement in Education, 1964.*

[5]*Crowder, Norman A., "Intrinsically Programmed Teaching Devices," 1959 INVITATIONAL CONFERENCE ON TESTING PROBLEMS, p. 40 ff., October 31, 1959.*

[6]*Committee on Measurement and Evaluation of the American Council on Education COLLEGE TESTING, p. 125, American Council on Education, Washington, D.C., 1959.*

measurements, a test of 90 questions was developed in connection with Cooperative Research Project No. 702, EFFECTS ON USE OF TESTS BY TEACHER TRAINED IN A SUMMER INSTITUTE.[1] Also at present, Robert L. Ebel is developing a group of measurement questions under the sponsorship of the National Council on Measurements Used in Education.

A variety of systems have been developed for classification of items. Unfortunately, many of these have proven

> ". . . too abstract, too general, and too ambitiously inclusive to help the test constructor very much in defining the content to be included and the abilities to be covered in his test."[2]

One of the more useful systems of classification may be found in the well-known TAXONOMY OF EDUCATIONAL OBJECTIVES.[3] The "Taxonomy" provides for classification of educational goals and test items in the cognitive areas under the following six major categories with a number of subcategories:

1. Knowledge (Remembering facts, terms, and principles in the form that they were learned)
2. Comprehension (Understanding material studies without necessarily relating it to other material)
3. Application (Using generalizations or other abstractions appropriately in concrete situations)
4. Analysis
5. Synthesis
6. Evaluation (Judging the value of material for a specified purpose)

Robert M. O'Brien at Northeastern University[4] has developed a program for item analysis of item responses on mark sense answer cards using a 1620 computer. However, the writers were unable to find any published reports of the effectiveness of using item analysis programs for improving learning.

A search of the literature does not reveal any instance of the use of IBM cards and accounting machines for convenient reproduction of carefully planned classroom tests.

---

[1] *Hastings, J. Thomas, Donald G. Beane and Emmanuel A. Lask, EFFECTS ON USE OF TESTS BY TEACHERS TRAINED IN A SUMMER INSTITUTE, 2:1 ff, Bureau of Educational Research, University of Illinois, Urbana, Illinois, 1961.*

[2] *Ebel, Robert L. and Dora E. Damrin, "Tests and Examinations," ENCYCLOPEDIA OF EDUCATIONAL RESEARCH, p. 1508, Third Edition, The Macmillan Company, New York, 1960.*

[3] *Bloom, Benjamin S., TAXONOMY OF EDUCATIONAL OBJECTIVES, Handbook I. Cognitive Domain, David McKay Company, New York, 1956. A similar volume listing a taxonomy for objectives in the affective domain has now been completed.*

[4] *This program is available through the IBM 1620 Users Group Library. Program number and title are 13.0.003, Northeastern University Test Scoring Program.*

PROCEDURES

The Sample

Students in two classes in Educational Tests and Measurements at North Dakota State University were used as the samples. This course was selected because of the convenient fact that the principal investigator taught it. The reference group consisted of 64 students registering for the class in the winter quarter (1963-64) and the experimental group, 81 students registering for the class in the spring quarter (1964). The class had three one-hour sessions per week for ten weeks—a total of thirty meetings, not including the final examination. All the students used as subjects attended at least 27 of the 30 class sessions.

Table 1 indicates the distribution by classes, colleges, and sex for the reference and experimental groups. As may be seen, the groups are reasonably similar in the characteristics listed. However, they were not selected on a strictly random basis as assumed in some of the statistical tests which follow.

TABLE 1

Characteristics of Groups According to Class, College, and Sex

Class			College			Sex		
	Reference*	Experimental**		Reference	Experimental		Reference	Experimental
Sophomore	4	2	Arts and			Male	30	44
Junior	19	40	Sciences	57	75	Female	34	37
Senior	36	33	Chemistry	1	2			
Grad.	5	6	Engineering	1	0			
	64	81	Home Economics	5	2			
			Agriculture	0	2			
				64	81			

*Reference—Winter quarter 63-64
**Experimental—Spring quarter '64

The School and College Ability Test (SCAT), Form 1A, was administered to both the reference and experimental groups. The mean scores were significantly different at the .01 level with the scores of the experimental group the higher. Table 2 indicates the difference in SCAT scores between the reference and experimental group.

TABLE 2

Comparison of The Means and Standard Deviations of Experimental and Reference Groups on SCAT

	Experimental	Reference
$\overline{X}$	83.64	77.30
S.D.	13.29	15.33
N	81	64
t	2.65**	

** $p < .01$

The General Plan

Approximately twelve hundred items were written and coded. These items were then pre-tested using as subjects students in the fall quarter, 1963, tests and measurements class. On the basis of item analysis and comments of students in the class, 930 items were found to be satisfactory for inclusion in the stockpile of items.

Students who took the course during the winter quarter 1963-64 constituted the reference group. Five unit tests and a final examination were constructed from the pool of items and administered to the reference group. The reference group was taught by the regular teaching methods.

Students who took the course during the spring quarter, 1964, constituted the experimental group. They were administered the same five unit tests and final examination as were the students in the reference group. However, in the experimental group, the students took a parallel practice test followed by discussion of items found difficult on the basis of item analysis. The practice test preceded the unit test by two class periods so as to permit adequate time for discussion of results.

Comparison was to be made of test scores achieved by the reference and experimental groups in an attempt to measure the effects of instruction based on the detailed analysis of practice test results.

Constructing the Tests

An analysis of course content was made as a basis for classification of test items. In addition to subject content, items were classified according to certain characteristics relating to the general skills measured.

The major divisions of the coding system are shown in Table 3. Subclassifications of the principal categories are given in Appendix A.

TABLE 3

Categories Within Classification System

Principal Subject Classifications (For subclassifications see Appendix A)

0. Scoring Systems
1. Regression and Correlation
2. Evaluating a Test—Validity and Reliability
3. Writing a Classroom Test
4. Using Teacher-Made Tests
5. Standardized "Ability" Tests
6. Standardized "Achievement" Tests
7. Interest Inventories
8. Measuring Personality
9. History of and Current Practices in Testing

Basic Skill Classification

1. Factual Knowledge
2. Application of Knowledge
3. Evaluation of Practices

Secondary Skill Classification

1. Verbal Reasoning
2. Statistical or Computational
3. Both Verbal and Computational

Item Number

Series Number (for use with several items based on the same options)

Special Criteria*

1. Negatively Stated
2. Used Charts or Tables
3. Correct Answers Seem Wrong
4. Question of Sufficiency of Data
5. Extraneous Data

*If an item were to possess more than one of the special criteria, then the item would be coded by the lowest number in the list of special criteria.

In order to insure that the test would have sufficient ceiling for measuring possible difference between reference and experimental groups, it was decided to use as many difficult items in the test as possible.

This meant that the content of the tests was mainly limited to the kinds of points students find hard to understand.

The selection of the items was made by the secretarial staff on the basis of the coded and statistical characteristics of the items. Since selection of items was made on an arbitrary basis by others, the investigator was not aware of which items were in the various tests. This procedure was developed to prevent the investigator from "teaching for the test" in a way that might have confounded results.

Forty items were selected for each unit test and eighty items for the final examination. Major content covered by each unit test is as follows:

Test 1—Scoring Systems
Regression Analysis and Correlation
Test 2—Evaluating a Test—Validity and Reliability
Test 3—Writing a Classroom Test
Using Teacher-Made Tests
Test 4—Standardized "Ability" Tests
Standardized "Achievement" Tests
Test 5—Interest Inventories
Measuring Personality
History of and Current Practices in Testing

In general, the tests were planned so that the average score would be between 50 and 60 percent correct. The practice tests were assembled so as to be parallel with the unit tests in coded characteristics and level of difficulty as far as possible. The specific characteristics of the test as they were assembled are indicated in Appendices B, C, and D.

The reliability coefficients for the unit tests and final examination based on the Kuder-Richardson Formula 21 are as follows:

Test 1	$r = .76$
Test 2	$r = .64$
Test 3	$r = .46$
Test 4	$r = .59$
Test 5	$r = .56$
Final Examination	$r = .79$
Sum of scores on all five unit tests	$r = .88$
Sum of scores on five unit tests and final examination	$r = .92$

In Appendix I, a number of illustrative test items are given which are fairly typical of those in the tests themselves.

Teaching Methods

Basically two teaching methods were used—one the regular method (reference group), the other the experimental.

In the regular method the lecture approach was used. Five lectures were given over a period of two weeks followed by a unit test. Students were permitted to ask questions during lectures but generally did not take much advantage of this opportunity.

In the reference group, students were given an immediate feedback of correct answers as soon as they had handed in answer sheets for the unit test. This was accomplished by giving the students a key of correct answers and allowing about five or ten minutes for each one to look over his test booklet before it was collected.

Students in the reference group were encouraged to receive individual tutoring on points that were difficult for them to understand. They were also permitted, for purposes of review, to study all of the unit test booklets in the instructor's office before the final examination. Seventeen of the sixty-one students in the class chose to receive individual assistance at various times during the course and also studied the unit booklets before examination. Of these, six were boys and eleven girls. (It may be recalled that the proportion of boys to girls in the total class was approximately equal).

In the experimental group, lectures were more abbreviated and compressed than in the reference group. Materials which were covered in five lectures in the reference group were covered in three in the experimental group.

Thus, the experimental group received the information at a much faster rate than did the reference group. It should be mentioned that five lectures were hardly sufficient in time to cover the materials of one unit of study for the course so that the experimental group with only three lectures was indeed under pressure to "absorb" information at a rapid rate.

On the fourth meeting of the experimental group in each unit, a test was given which will be designated as a "practice" test—although it actually counted half as much as the unit test in determining course grades. The practice test closely paralleled the unit test in content and difficulty. As soon as students had handed in their answer sheets, they were given a key of correct answers and they were permitted to keep the practice test booklets for study.

On the fifth meeting, the instructor explained certain items on the practice test which many in the group had missed as shown by the item analysis of the practice test. There was an opportunity for members of the class to raise questions and discuss any points which they found confusing. Although the experimental group was quite large (81 students), there was considerable participation by its members in discussion of some of the points covered on the practice tests.

Then, as in the case of the reference group, the experimental group took the unit test on the sixth meeting. However, in contrast to the reference group, they were permitted to keep copies of the unit test for their own review as they previously had been permitted to keep copies of the practice test.

As was true of the reference group, members of the experimental group were encouraged to receive individual tutoring. Twenty of the group availed themselves of this opportunity, thirteen boys and seven girls. A summary of the individual assistance provided to both reference and experimental groups is given in Appendix E.

Both the reference and experimental groups used the same tests and measurements textbook for reference. However, much of the material covered in the unit tests and final examination was based on class lecture and discussion.

Comparisons of Reference and Experimental Classes Considered as a Whole

Mean scores and variances are given for the unit tests and the final examination in Table 4.

Table 4

Means and Variances of Test Scores of the Reference, Experimental, and Combined Groups

Test Number	Means			Variances	
	Reference (N=64)	Experimental (N=81)	Total Group	Reference	Experimental
1	19.766	22.198	20.982	25.357	41.611
2	18.422	20.901	19.661	25.899	27.965
3	25.922	28.161	27.041	20.264	15.211
4	24.219	26.185	25.202	23.221	22.127
5	18.891	19.938	19.414	18.353	22.509
Final*	20.938	23.148	22.043	18.710	23.215
All Tests Including Final	21.359	23.421	22.390		

*Final based on ½ number of correct items for purposes of easier comparison with the unit test means.

Inspection reveals that differences among the variances in Table 4 are not greater than might be reasonably expected by chance. Accordingly, main attention was directed to possible variation in mean scores between the reference and experimental groups. A 2 x 6 analysis of variance was computed with the results as indicated in Table 5.

TABLE 5

Analysis of Variance on Unit Tests and Final Examination for Total Reference and Experimental Groups

Source	d.f.	Sum of Squares	Mean Squares	F
Between Subjects	144	14069.719		
Instruction	1	912.208	912.208	9.915***
error (b)	143	13157.511	92.010	
Within subjects	725	14392.840		
Tests	5	6984.072	1396.814	135.744***
Tests x Instruction	5	50.40	10.08	.979
error (w)	715	7358.368	10.29	
Total	869	28462.559		

***Significant at .01 level

As indicated in Table 5, differences between scores obtained by the reference and experimental group were significant at the .01 level with the experimental group making the higher scores. Also the difference among the mean scores for the five unit tests and the final tests was significant at the .01 level.

Using Tukey's test, it was ascertained that the combined means for the unit tests and the final examination were all significantly different one from another at the .01 level with the exception of the means for Test 2 and 5 which were almost the same.

No significant interaction effects were found.

On the basis of the analysis of variance, then, it appears that the experimental group achieved better overall on the six tests than did the reference group. However, the experimental group was of higher general ability than the reference group as measured by the School and College Ability Test (Table 2). Therefore, differences between the two groups could possibly be ascribed to initial differences in ability level rather than to differences between teaching methods. Accordingly, an analysis of covariance, with the School and College Ability Test (SCAT) serving as an adjustment variable, was conducted for each of the individual unit tests, the final examination, and the sum of all six tests. Results of these co-variant adjustments together with analysis of variance on the adjusted data are presented in Table 6 on the following page.

TABLE 6

Summary of Analysis of Variance and Covariance of Test Scores for the Total Group

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Ref.	Exp.	F	Ref.	Exp.	Ref.	Exp.	F	Ref.	Exp.	F
Test 1	20.50	22.20	6.14**	25.36	41.61	.50	.52	2.32	20.50	21.22	1.64
Test 2	18.42	20.90	8.12***	25.90	27.97	.41	.48	1.08	19.00	20.45	3.27*
Test 3	25.92	28.16	10.27***	20.26	15.21	.57	.60	.05	26.53	27.68	3.92**
Test 4	24.22	26.19	6.11**	23.22	22.13	.39	.45	.50	24.72	25.79	2.10
Test 5	18.89	19.94	1.90	18.35	22.51	.45	.33	.03	19.32	19.60	.15
Final	41.77	46.30	8.23***	74.84	92.86	.40	.59	4.76**	43.02	45.39	3.00*
Sum of Tests 1-5	107.22	117.27	9.47***	349.09	407.05	.59	.59	.83	110.03	115.05	3.39*
Sum of All Tests	149.09	163.57	9.97***	654.20	824.95	.56	.61	2.05	153.06	160.43	3.71*
SCAT	77.30	83.64	7.03***	235.24	176.55						

NUMBER OF STUDENTS:
Reference—64
Experimental—81
Total—145

*—Significant at .10 level
**—Significant at .05 level
***—Significant at .01 level

It may be noted that the total adjusted mean score on the five unit tests and final examination of the experimental group is higher than that of the reference group. The difference is significant at the .10 level. Thus, generally speaking, the experimental group remains higher even after the means have been adjusted. However, it should be noted that in Test 5 (covering interest inventories, personality measurement and history and current practices in testing), there is substantially no difference between the adjusted mean scores for the reference group and the experimental group.

Differences in variance for the various groups of test scores were not high and with the exception of the final test, the regression lines used in the analysis of co-variance are reasonably parallel. In the final test, the regression lines were found to be nonparallel at the .05 level of significance.

Percent Correct on Individual Items by Group

Differences between reference and experimental group for percent correct on individual items were noted. Table 7 indicates the number of items by test for which the difference in the percents correct was significant at the .05 level.

Table 7

Number of Items Per Test on Which Reference and Experimental Differ at the .05 Level in Percent of Correct Responses*

	Reference High	Experimental High
Test 1	3	7
Test 2	3	7
Test 3	3	12
Test 4	4	8
Test 5	5	6
Final	7	13
Total	25	53

*As indicated by the data on some items the reference group scored much higher than the experimental and vice versa.

As may be noted, when there were considerable differences in percent correct on individual items between reference and experimental groups, it was generally the experimental group which was the higher.

Comparative Difficulty of Practice and Unit Tests

Some comment should be made concerning the practice tests taken by the experimental group. As might be expected, the experimental group found the practice tests more difficult than the unit tests. Means for the unit tests were all higher than the corresponding practice tests. All these differences reached significance at the .01 level with the exception of unit test five, which was at the .20 level. A summary of the differences between mean scores on the practice and unit tests are given in the Appendix F.

Intercorrelations

A summary of intercorrelations among test scores on the unit tests and final examination for the reference and the experimental groups is given in Appendix G. These intercorrelations ranged from .37 to .75.

Scores Obtained by Students with High and Low Ability as Measured by SCAT

Analysis was made of scores obtained by students in the upper fourth on SCAT in the reference and experimental groups. Similarly, an analysis was made comparing reference and experimental students falling in the lower fourth on SCAT in their respective groups. The results of these analyses are presented in Tables 8 and 9.

For the students scoring high on the SCAT, there is a considerable difference in the adjusted mean in favor of the experimental group. However, for the lower one-fourth on SCAT, the difference, while still in favor of the experimental, is less. Of particular note is the fact that the adjusted mean on the final examination is approximately nine points higher for the top one-fourth of the experimental compared to the reference group; whereas for the bottom one-fourth, the experimental is only about a point higher on the final examination.

As an additional comment, it should be observed that the division of men and women in each of the subgroups is approximately equal with the exception of the lower one-fourth on the reference, where the women are approximately double in number to the men.

Scores Obtained by Men and Women Considered Separately

Tables 10 and 11 present analysis of variance and co-variance results for men and women considered separately.

Results for the reference and experimental men's groups (with the exception of Unit Test 5) show a strong trend in the direction of the experimental group scoring the higher. It should be noted that the F indicating non-parallel regression lines is significant at the .05 level for Unit Test 3.

TABLE 8

Summary of Analysis of Variance and Covariance Results for Upper One-Fourth on SCAT

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Ref.	Exp.	F	Ref.	Exp.	Ref.	Exp.	F	Ref.	Exp.	F
Test 1	23.12	27.10	7.24**	17.72	20.73	.46	.34	.12	23.81	26.56	3.48*
Test 2	20.06	23.90	5.01**	33.53	20.31	.39	.25	.53	20.70	23.39	2.34
Test 3	28.15	31.25	6.34**	21.05	7.88	.37	.55	.06	28.77	30.74	2.65
Test 4	26.19	30.05	6.54**	25.90	15.84	.48	.45	.32	27.01	29.39	2.68
Test 5	21.56	23.05	1.00	18.53	20.68	.28	.53	.59	22.32	22.44	.007
Final	45.25	55.00	14.08***	71.67	50.84	.46	.24	.86	46.31	54.15	8.82***
Sum of Tests 1-5	119.06	135.35	9.11***	05.93	221.50	.55	.46	.41	122.04	132.97	4.71**
Sum of All Tests	164.31	190.35	11.91***	02.36	429.61	.55	.41	.63	168.57	186.94	6.53**
SCAT	95.81	99.50	4.61**	23.12	26.45						

NUMBER OF STUDENTS:
Reference—16 (Men—7; Women—9)
Experimental—20 (Men—10; Women—10)
Total—36

*—Significant at .10 level
**—Significant at .05 level
***—Significant at .01 level

TABLE 9

Summary of Analysis of Variance and Covariance Results for Lower One-Fourth on SCAT

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Ref.	Exp.	F	Ref.	Exp.	Ref.	Exp.	F	Ref.	Exp.	F
Test 1	16.37	18.75	1.26	17.98	56.93	.23	.43	2.22	16.90	18.33	.35
Test 2	15.37	17.20	1.57	20.25	17.75	.24	.15	.00	15.85	16.82	.35
Test 3	22.12	25.10	6.06**	16.12	10.52	.24	.29	.07	22.66	24.67	2.29
Test 4	21.25	24.40	5.24**	17.93	15.95	.38	.14	.20	20.91	24.67	5.92**
Test 5	16.19	18.60	3.51*	4.83	22.57	.00	.16	.01	15.99	18.76	3.51*
Final	36.87	39.25	.59	82.65	87.99	.04	.21	.37	37.49	38.76	.13
Sum of Tests 1-5	91.31	104.05	5.55**	178.36	322.05	.04	.26	.81	92.31	103.25	3.18*
Sum of All Tests	128.19	143.30	3.57*	421.50	684.43	.01	.25	.04	129.80	142.01	1.81
SCAT	57.00	65.20	9.17***	87.62	40.66						

NUMBER OF STUDENTS:
Reference—16 (Men—5; Women—11)
Experimental—20 (Men—10; Women—10)
Total—36

*—Significant at .10 level
**—Significant at .05 level
***—Significant at .01 level

TABLE 10

Summary of Analysis of Variance and Covariance for Men

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Ref.	Exp.	F	Ref.	Exp.	Ref.	Exp.	F	Ref.	Exp.	F
Test 1	19.90	22.64	4.68**	19.61	34.52	.55	.20	.57	20.31	22.35	2.75
Test 2	18.20	20.93	5.52**	19.68	27.13	.37	.31	.08	18.59	20.65	3.38*
Test 3	25.13	28.05	10.42***	19.29	11.30	.73	.33	4.84**	25.63	27.70	7.11***
Test 4	23.77	25.89	4.52**	17.22	18.10	.55	.25	.91	24.16	25.62	2.38
Test 5	18.40	19.30	.78	20.94	16.49	.36	.02	.43	18.57	19.18	.35
Final	40.37	46.86	10.79***	79.00	63.52	.54	.25	1.33	41.14	46.33	7.68***
Sum of Tests 1-5	125.40	116.82	8.27***	289.77	275.50	.66	.32	1.41	107.32	115.51	5.21**
Sum of All Tests	145.77	163.68	10.04***	613.70	540.73	.65	.31	1.56	148.46	161.84	6.81**
SCAT	78.40	83.77	2.91*	215.70	141.99						

NUMBER OF STUDENTS:
Reference—30
Experimental—44
Total—74

*—Significant at .10 level
**—Significant at .05 level
***—Significant at .01 level

TABLE 11

Summary of Analysis of Variance and Covariance for Women

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Ref.	Exp.	F	Ref.	Exp.	Ref.	Exp.	F	Ref.	Exp.	F
Test 1	19.65	21.68	1.76	31.14	50.73	.48	.78	7.95***	20.65	20.76	.61
Test 2	18.62	20.86	2.90*	32.06	29.73	.44	.65	1.21	19.35	20.20	.50
Test 3	26.62	28.30	2.44	20.67	20.27	.48	.80	3.74*	27.32	27.65	.16
Test 4	24.62	26.54	2.34	28.85	27.31	.31	.61	2.32	25.20	26.00	.48
Test 5	19.32	20.70	1.47	16.23	29.21	.57	.59	1.32	19.98	20.10	.01
Final	43.21	45.62	1.02	69.56	129.56	.31	.82	16.87***	44.66	44.28	.04
Sum of Tests 1-5	108.82	117.81	2.90*	406.15	574.94	.56	.77	4.27**	112.37	114.55	.29
Sum of All Tests	152.03	163.43	2.42	696.39	1187.31	.53	.80	8.52***	157.07	158.80	.10
SCAT	76.32	83.49	-----	248.99	216.85						

NUMBER OF STUDENTS:
Reference—34
Experimental—37
Total—71

*—Significant at .10 level
**—Significant at .05 level
***—Significant at .01 level

In the case of the analyses of the women's scores, the trend toward higher scores for the experimental is much less pronounced but with only one test score in favor of the reference group. (Of the 48 analyses of co-variance in this study, this was the only one where the reference group scored higher.) The fact that the F's indicating nonparallel regression lines are high makes it difficult to interpret the meaning of the covariant adjustment. Of some interest is the fact that the correlation between SCAT and criterion scores is very high for the women's experimental group with r's obtain for the final examination, the sum of the five unit tests, and the total sum of all scores, computed as .82, .77, and .81, respectively.

Some further analysis was made in which the upper one-half of men on the SCAT for both reference and experimental groups were compared with one another. A similar analysis was conducted for low scoring men on SCAT, high scoring women, and low scoring women. Generally speaking, although the results were always in the direction of the experimental group obtaining higher scores, the differences between the adjustment mean did not reach significance at the .10 level **with one exception**—low scoring men. Results of the analysis on the low scoring men are given in Table 12.

As may be observed, differences between reference and experimental groups were significant at the .01 level for the final examination, the sum of Tests 1 through 5, and the total sum of all scores.

Relative Success of Men and Women Under the Two Teaching Conditions

One final point was considered concerning the relative success of men and women under the regular and experimental teaching condition. An analysis was made of the differences between men and women on the criterion test within the reference and experimental groups considered separately. A summary of these analyses is given in Appendix H.

Generally speaking, the women did better than the men under the regular conditions.

Results were different for the experimental group. There the means for men and women were approximately the same for all criterion tests. Correlation with SCAT test scores was very low for the men and high for the women. The regression lines for the analysis of covariance were nonparallel as evidenced by the high F's obtained.

Student Attitude Toward Practice Tests

It appeared that the students in both the reference and the experimental groups favored the idea of practice tests as used in this project. Table 13 indicates the students' answers to the question, "Would you like to take practice tests followed by discussion even though this would reduce lecture time?"

TABLE 12

Summary of Analysis of Variance and Covariance for Lower Half of Men on SCAT Scores

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Ref.	Exp.	F	Ref.	Exp.	Ref.	Exp.	F	Ref.	Exp.	F
Test 1	17.53	22.32	6.04**	18.41	44.04	.26	.16	.01	17.97	22.02	3.86*
Test 2	16.60	20.32	4.52**	16.40	34.51	.24	.41	1.34	17.31	19.83	2.01
Test 3	22.53	27.73	17.99***	19.27	9.45	.69	.45	1.06	23.43	27.18	11.99***
Test 4	21.13	25.64	14.56***	11.84	12.88	.11	.19	.14	21.35	25.49	10.84***
Test 5	16.67	19.59	4.23**	14.67	20.25	.12	.17	.09	16.60	19.64	3.96*
Final	35.60	46.14	13.33***	68.54	78.12	.31	.20	.00	36.48	45.54	9.06***
Sum of Tests 1-5	94.47	115.18	14.65***	235.70	278.25	.39	.26	.00	96.56	113.75	9.65***
Sum of All Tests	130.07	161.32	15.72***	494.64	593.56	.38	.25	.00	133.04	159.29	10.56***
SCAT	67.27	74.41	4.36	134.82	75.95						

NUMBER OF STUDENTS:
Reference—15
Experimental—22
Total—37

*—Significant at .10 level
**—Significant at .05 level
***—Significant at .01 level

This question was asked both the first and last days of the course. It may be recalled that the practice tests actually were used in determining student grades to the extent of counting half as much as the unit tests.

Table 13

Teaching Method Preferred by Students at Beginning and End of Course

	L* — L	L — T**	T — L	T — T
Experimental.....	2 (3%)	17 (21%)	9 (11%)	52 (65%)
Reference	11 (19%)	18 (32%)	6 (11%)	22 (38%)

*L = preference for lecture method
**T = preference for practice test method

First letter indicates preference at beginning of course; second letter indicates preference at end of course. Thus, "L-L" means student preferred lecture method both at beginning and end of course.

Total Preferring Practice Test Method at End of Course

Experimental — 69 (86%)
Reference — 40 (70%)

DISCUSSION OF RESULTS

Results for Total Group

In the present research, the reference group taught by the regular lecture method, enjoyed certain advantages which might have been expected to balance out the effects of practice testing as employed in the experimental group. Members of the reference group received more individual tutoring than did the students in the experimental group. (See Appendix E). Furthermore, after planning the lectures given to the reference group, it was possible to cover in class most of the facts and concepts required for answering test questions. On the other hand, in the experimental group, a number of points were not covered in lectures nor were they discussed in connection with the practice testing, simply because time was at such a premium.

It should be noted that "Introductory Tests and Measurements," as presented at North Dakota State University, is a course covering considerable content so that even with the usual lecture method of teaching, it is not easy to give adequate consideration to all the matters which should be tested. Some members of the experimental group commented that they had never in their college experience been given so much information in lecture form within such a short length of time. (It may be recalled that they were given three lectures covering what had been presented to the reference group in five lectures.) The instructor himself feels that for the experimental group, the lectures were compressed to the extent that it would be difficult for many students to keep up with the flow of information.

It should be added also that in the reference group, contrary to common practice in many lecture and discussion type classes, the students were given a quick feed-back of results of answers to individual items and they were permitted to study the five unit tests in the instructor's office prior to the final examination.

Despite the advantages mentioned, the experimental group clearly performed better on the criterion tests than did the reference group. After adjustment for differences on SCAT scores, 48 comparisons were made between experimental students and reference students (See Tables 6, 8, 9, 10, 11 and 12.) While not all differences were statistically significant, 47 of the 48 were in favor of students in the experimental group. This trend strongly suggests that the

method by which the experimental students were taught does produce better learning which is of greater or lesser importance depending on additional factors such as sex, specific unit content, and ability of students.

Results for Subgroups

Examining results for subgroups, it appears that both high and low scoring persons on the SCAT benefit from the effect of practice testing. However, the high persons seem to profit more, particularly as measured by achievement on the final examination. It also appears that men benefited to a greater degree than women from practice testing, and especially those men scoring low on the test of general ability (SCAT).

It is not possible to give a definite answer as to why the men seem to benefit more than the women from the practice testing method, but the following information may be pertinent. As indicated in Appendix E, the women making low scores on SCAT took more advantage of the opportunity for individual tutoring than did the men. Contrariwise, men with low scores on SCAT took the greater advantage of tutoring opportunities in the experimental group. Thus, it is possible that in the reference group women's scores were upgraded and in the experimental group men's scores were increased because of the effects of individual tutoring.

Likely there are other factors which may help explain the difference in achievement between men and women under the two conditions. Possibly women at North Dakota State University, as a group, are more highly motivated than men, and hence better able to adjust to variation in teaching methods. However, the exact nature of factors which may result in different achievement by men and women are beyond the scope of this study.

Difference in Percent Correct for Reference and Experimental Groups on Individual Items

As was indicated in Table 7, of the 240 items on the unit and final tests, there were 78 items on which there was a difference significant at the .05 level in percent correct between reference and experimental groups. On 53 items, the experimental scored higher and on 25 the reference scored higher.

The present experiment was not designed to provide an explanation for these differences in success on individual items between the reference and experimental group. However, it would seem reasonable to suggest four possible general explanations for the differences.

1. Variation in informational background from differences in previous course work.

2. Variation in knowledge from difference in the lecture presentation between the two groups.
3. Variation in knowledge relating to the effects of administering practice tests to one group but not the other.
4. Effects of random guessing (particularly on the more difficult test items).

Considering the overall results of this experiment, it would seem reasonable to believe that the third explanation, the effects of practice testing, was a significant factor in determining the number of correct responses to individual items. As previously noted, the experimental group which had the advantages of the practice test, was significantly higher at the .05 level for approximately twice as many items as was the reference group. (See Table 7.)

Samples of these items may be found in Appendix I.

Difficulty of the Unit Tests

As indicated in Table 4, there was variation in difficulty among the various unit tests and the final examination. Generally speaking, tests on units relating to applied aspects of measurement (Unit Tests 3 and 4) were easier for both the reference and experimental groups where mean scores were six or seven points higher (forty item tests) than mean scores on the other tests.

Special mention should be made of Test 5 which covered a unit on Personality Testing and History and Current Trends in Testing. Of all the units covered, this one was most comprehensive. It was the one unit where outside reading was particularly necessary to answer many of the test questions, and the reading assignments were extensive, consisting of more than 150 pages of somewhat technical material.

The practice test for Unit 5 was less parallel to the unit test than were the other practice tests. This is illustrated in Appendix B. As may be recalled, selection of test items for the practice tests was accomplished in an arbitrary manner using a system akin to a stratified random selection procedure. Thus, given a considerable diversity of content in the unit of study, it was not to be expected that the practice test would closely parallel the unit test.

It should also be noted that Unit 5 test was given at the end of the quarter when students were under pressure in other classes to finish papers and take tests. As a result, it is probable that many were lax in completing the assigned reading.

Scores on Unit Test 5 were generally low for both reference and experimental groups. In this case, the lack of student preparation may have had more effect upon test results than whether or not a practice test had been given.

Attitude of Students Toward Practice Testing

The students in this experiment indicated that they favored being instructed through use of practice tests (See Table 13). This finding is consistent with results of other studies mentioned in Related Research.

The instructor believes that the students like practice tests because the items in such tests help to define for students what they are supposed to know. Furthermore, students appear to relish the opportunity to challenge the correctness of the instructor's key of answers. The privilege of retaining practice tests as well as the unit tests for review purposes is evidently also well liked by students. In addition, they seemed to appreciate the quick feed-back to them of correct responses to individual items. Many students remained after class in order to continue discussion of various items on the practice tests.

It should be pointed out that in this particular research, the experimental group was under considerable pressure (as to a lesser degree was the reference group) because more material was covered than ordinarily should be covered in a ten weeks unit of study. (At least in the instructor's opinion this is so.) It was felt that the differences in results from the two teaching methods might be more apparent where the students were put in a very challenging situation than would have been the case had a normal amount of material been studied. Thus, one might possibly have expected the experimental group to have resented the practice tests as causing lectures to be abbreviated, compressed, and consequently more difficult to follow. However, such was not the case. If anything, the experimental group favored the idea of practice testing more than did the reference group where the practice tests were not administered.

From the point of view of the instructor, the use of practice tests afforded a decided advantage in teaching. Through item analysis performed on the 1620 IBM computer, it was possible to identify items which students found difficult and to examine the pattern of responses to the various distractors. With this information, the instructor was able to plan the discussion so as to consider pertinent points relating to student misunderstandings **before the unit tests were given.**

USE OF DATA PROCESSING EQUIPMENT

Basically the data processing equipment was used in two ways:

1. To reproduce tests directly from the item files following certain coded specifications.
2. To analyze test results for instructional purposes.

Reproducing the Tests

It was found convenient and reasonably economical to reproduce the tests using the IBM punched cards. A slight disadvantage in the use of the cards is that a typical test item required six or seven cards. *The following data* indicates the costs in the present research of using IBM cards for reproduction of the tests. Costs will vary according to institutional opportunity for obtaining data processing equipment services.

Initial costs of punching cards for approximately 1,000 items:	
100 hours secretarial time @ $1.50 per hour	$150.00
100 hours 026 IBM Key Punch @ $.27 per hour	27.00
6,000 cards @ $1.03 per 1,000	6.18
Costs of reproducing 200 copies of 5 unit tests and the final (a total of 240 items):	
2 hours 407 IBM Accounting Machine for cutting stencils for mimeographing @ $3.19 per hour	6.36
20 reams of mimeograph paper @ $1.00 per ream	20.00
Estimated cost of mimeographing and assembly of test booklets	60.00
	$269.54

Test Scoring and Item Analysis

It was possible to schedule use of the 1620 IBM computer so that tests could be scored either the same day or the day following the day of test administration. Approximate cost of test scoring including item analysis are given as follows:

	4,000 cards @ $1.03 per 1,000	$ 4.12
2 hours	519 Reproducer with Mark Sense Attachment @ $2.00 per hour	4.00
4 hours	1620 IBM Computer @ $10.00 per hour	40.00
2 hours	407 IBM Accounting Machine @ $3.19 per hour	6.38
	Total	$54.50

Thus the total cost of (1) Reproducing the Tests and (2) Test Scoring and Item Analysis is $324.04.

The item file for the course in Introductory Tests and Measurements, punched on IBM cards, with an explanation of the coding system will be made available.

The principal author is presently engaged in arrangements with an outside testing agency with a view to their distributing these materials. If this does not work out, the author will then arrange for distribution through North Dakota State University.

In either case, the test materials will be distributed at cost.

SUMMARY

A major purpose of this experiment was to evaluate the use of practice tests as a learning aid. It was also hoped to demonstrate how data processing equipment could be used effectively to reproduce tests and analyze test results for improvement of instruction.

Initially, a class of students in tests and measurements at North Dakota State University (the reference group) was administered five unit tests and a final examination. The teaching procedures follow the usual lecture-discussion method.

A second class in tests and measurements (the experimental group) was taught by a different method but administered the same unit tests and final examination as the reference group. Lectures were compressed to permit time for a practice test followed by discussion to precede each unit test. The practice tests were parallel in content and difficulty with the unit tests.

Results of the study were as follows:

1. The data indicated higher mean scores for the group which was administered the practice tests using SCAT scores for a covariant adjustment.
2. Forty-seven out of 48 different comparisons favored the experimental group when separate consideration was given to various subgroups.
   a. Adjusted mean scores on the criterion tests for the upper one-fourth, based on SCAT, of the experimental group were higher than for the upper one-fourth, on SCAT, of the reference group. To a lesser degree the lower one-fourth of the experimental group scored higher than the corresponding reference group.
   b. Adjusted mean scores for men were higher on the experimental group than on the reference group. However, the differences between the women taught by the regular method and those taught by the experimental method using the practice tests, while consistently in favor of the experimental group, were small and generally not statistically significant.

3. A large majority of the students favored the use of practice tests followed by discussion of results.

4. The use of data processing equipment provided an effective means for reproducing test booklets following carefully planned specifications for test content. Subsequent item analysis of test results served as an aid for planning instruction to cover points not fully understood by the class.

Overall, use of practice tests and data processing equipment is believed to have contributed to effective learning in a class in tests and measurements at North Dakota State University.

APPENDIX A

Major Subject Classifications and Sub-Classifications
of the Item Coding System

0. Scoring Systems

1. Measurement Scales
   1. Labels
   2. Categories
   3. Ordinal
   4. Interval
   5. Ratio

2. Central Tendencies
   1. Mean
   2. Median
   3. Mode

3. Normal Distribution
   1. General characteristics
   2. Skewness (position of median and mode)

4. Dispersion, Spread or Position (Based on concept of normal distribution)
   1. Range
   2. Deviation scores
   3. Standard deviation (Including effects of adding, subtracting, etc.)

5. Percentiles

6. Standard Scores Types
   1. z and similar standard systems
   2. Deviation IQ scores (Wechsler, Stanford-Binet, Group Tests)
   3. Stanines
   4. Large scale testing programs using standard scores (CEEB, AGCT, ACT, National Merit, SCAT, STEP, etc.)

7. Grade Placement
8. MA/CA
9. Relationship among scoring systems (percentiles, standard scores, etc.)
10. Evaluation of Scoring Systems
    1. Percentiles
    2. Standard score, etc.
    3. Grade placement
    4. MA/CA

1. Regression and Correlation Concepts

1. Inspectional Techniques
    1. Lists and scores
    2. Contingency tables
    3. Scattergrams
    4. Freehand line of best fit
2. Mathematical Techniques (Regression)
    1. Conversion to z
    2. Least-squares solution
    3. Using and interpreting the "best-fit" line
    4. Distribution of scores around regression line, including homoscedasticity
3. Meaning of Correlation
    1. Definition
    2. Correlation coefficient and the best-fit lines
    3. Predicting scores from correlation
4. Standard Error of Estimate
    1. Formula and use
    2. Interpretation of
5. Problems in Interpreting and Use of Correlation Coefficient and Regression Equations
    1. Correlation and variability
    2. Improving prediction over chance
    3. Effectiveness of prediction related to level at which discrimination is desired
    4. Regression effect
    5. Satisfactory correlation for selection procedure
    6. Interpreting a series of intercorrelations
    7. Curvilinear correlation

2. Evaluating a Test

1. Predictive Validity
    1. Predictor compared with assessment tests
    2. Ways of measuring predictive validity
    3. Intercorrelations among tests
    4. Cross validation
2. Concurrent Validity
3. Content Validity
    1. Systematic analysis (Grid)

2. Taking the test
3. Factor analysis
4. Avoid judging test by title
5. Compared with face validity
6. Critical incidents technique

4. Construct Validity (Defined)
5. The Concept of Reliability
   1. Observed vs. true score (defined)
   2. Definition of reliability coefficient
   3. Sources of error in the test score
   4. Test-retest coefficient
   5. Split test coefficient (Spearman-Brown)—
      Item intercorrelation (Kuder-Richardson)
   6. Equivalent form
   7. Reliability coefficient as regression line (true score-observed score relationship)
   8. Standard error of measurement
6. Relationship of Reliability and Validity
   1. Comparison of terms and concepts
   2. Reliability affects validity
   3. Similarity of content validity and sampling reliability
   4. Correction for attenuation
   5. Effects of variability on validity and variability
7. Sources of Information (listings) About Tests
8. Evaluating the Test Manual
   1. Validity data
   2. Reliability data
   3. Scales and norms
   4. Instruction for administering and scoring
   5. Interpreting scores
   6. Information on extent to which test is speeded
   7. Appropriateness of difficulty
   8. Relationship between reliability and independence of measure

3. Writing the Classroom Test

1. Planning the Test
   1. Defining objectives
   2. Detailed plan
2. Writing Essay Questions
3. Scoring Essay-Type Questions
   1. Critical points and numerical weight
   2. Score all papers on same questions
   3. Separate into groups for further comparison
   4. Avoid personal bias in scoring
   5. Need for general scoring categories
   6. Need for written comments
4. Comparing Several Kinds of Objective-Type Questions
   1. Multiple choice
   2. True - false

3. Matching
4. Completion

5. Essay Compared to Objective-Type Questions
   1. Abilities measured
   2. Scope
   3. Incentive to pupils
   4. Ease of preparation
   5. Scoring

6. Writing the Test Item
   1. Include questions requiring application of learned facts and skills
   2. Stating questions clearly
   3. Proper English usage
   4. Avoiding specific determiners
   5. Avoiding questions on trivial points
   6. Using plausible distractors

7. Evaluating Test Questions
   1. Student comment
   2. Item analysis

8. Stockpiling of Questions
   1. Convenient source of questions
   2. Continuous procedure
   3. Control over test difficulty

4. Using Teacher-Made Tests for Instruction and Evaluation

1. Improvement of Instruction Through Testing
   1. Identifying areas where help is needed (item analysis)
   2. Discussion of troublesome test items
   3. Testing frequently

2. Purposes in Testing
   1. Testing achievement (parallel to classwork, moderate dispersion of scores)
   2. Testing to rank students (critical points of learning, high dispersion of scores)
   3. Testing for diagnostic purposes (subtests and scores)

3. Creating Desirable Attitudes of Pupils Toward Tests
   1. Preparation of pupils for taking test
   2. Prompt grading, return, and discussion of test

4. Special Problems
   1. Cheating
   2. Correction for guessing
   3. Grading on curve
   4. Appropriate per cent right as criterion for passing
   5. Relative emphasis on speed and power

5. Ability Tests

1. General Ability Tests As Measures of Intelligence
   1. Disagreement on meaning of intelligence
   2. Content of test a reflection of author's definition

3. Variation in score according to test taken
4. Effects of environment on intelligence scores

2. Individual
   1. Defined
   2. Advantages (uses)
   3. Disadvantages
   4. Examples of
3. Group Tests
   1. Defined
   2. Advantages (uses)
   3. Disadvantages
   4. Examples of
4. Confusion About the Meaning of General Ability Scores
   1. Names of tests which measure about the same thing
   2. "What does the term 'I.Q.' mean?"
   3. Avoid the term I.Q.
   4. "Culture-free" test (fairness to cultural groups)
5. Breakdown of General into Separate Abilities
   1. Verbal—quantitative
   2. Verbal—performance (non-verbal)
   3. Factor analysis (advantages and disadvantages)
   4. Example of tests based on factor analysis (GATB)
6. Confusion about the Meaning of "Aptitude" Testing
7. Special Aptitudes
   1. Examples of tests
   2. Evaluation of tests

6. Standardized Achievement Tests

1. Kinds of Achievement Tests
   1. Batteries
   2. Diagnostic and survey tests
   3. Specific subject matter tests
2. Local Norms vs. National Norms
   1. Appropriateness of national
   2. Appropriateness of local
   3. Determining a "passing score"
3. Interpretation of Scores
   1. Effect of socio-economic condition
   2. Achievement tests as measures of demonstrated intelligence
   3. Relationship between teaching objectives and test content
   4. Reading a test profile
   5. Comparing the scores made by different individuals
   6. Measuring growth (problems of regression effect and unreliability)
   7. Group norms compared with individual norms
4. Uses of Standardized Achievement Tests
   1. Recognition of general level and variability of class achievement
   2. Identification of areas where classes as a whole needs special help (item analysis and early testing)

3. Identification of students with special strengths and weaknesses (diagnosis)
4. Discussion of answers to test questions as learning exercises
5. Placement for homogeneous grouping
6. Broadening teacher objectives through consideration and evaluation of published tests

5. Developing a School Testing Program
   1. Involving school personnel (teachers)
   2. Establishing testing needs
   3. Training teachers in use of test scores
   4. Interpreting scores to parents and pupils
   5. An ideal program

7. Interest Inventories

1. Information About the "Kuder"
   1. Rationale of the tests
   2. Interest areas used
   3. Scoring systems
   4. Interpretation
2. Information About the "Strong"
   1. Rationale of the tests
   2. Interest areas used
   3. Scoring systems
   4. Interpretation
3. Possible Appropriate Uses
   1. To call attention to unrealistic personal preferences
   2. To arouse interest in several occupations
4. Weaknesses Because of Invalid Answers
   1. Faking answers
   2. Questions sometimes beyond the experience of the person taking the test
   3. Interest not stabilized or may change with experience
   4. Answers ignore capabilities of the test taker
5. Difficulties in Interpreting Scores
   1. Scores not an indication of the intensity of interest
   2. Interest areas and existing occupations not always clearly related
   3. Results of different inventories are not always in agreement
   4. Confusion in normative scale (Ipsative—Normative)
   5. Rationale of some tests says the typical member of an occupation is the desirable member
   6. Scores may sometime be misinterpreted as measures of aptitude
   7. Interest inventories possibly not as valid as other information
6. Opinions and Attitudes

8. Measurement of Personality

1. Definitions of Personality
2. Problem Check-Lists
3. Personality Inventories
   1. Problems of getting truthful responses

2. The forced choice technique
3. Possible misinterpretation of results

4. The MMPI (A Personality Inventory)
   1. Scoring categories
   2. "Validity scores"
   3. Interpreting results of

5. Projective Techniques
   1. Rationale
   2. Examples (Sentence completion, Thematic Apperception Test, Rorschach)
   3. Scoring systems
   4. Validity and reliability

6. Observation of Behaviors
   1. Ratings
   2. Situational

9. History and Current Practice in Testing

1. Early Developments in Testing
   1. Psycho-physicists
   2. Binet and others
   3. Oral testing and essay testing

2. Development of Group Testing
   1. Army Alpha, Army Beta
   2. Rationale (speed and facts)
   3. Rapid growth of group tests
   4. Acceptance of personality and interest measures as valid instruments

3. Recent Developments in Testing
   1. Emphasis on power and reasoning type tests
   2. Trend toward achievement testing over intelligence testing
   3. Disillusionment with validity of tests of interest, personality and creativity
   4. Growth of testing for selection purposes
   5. Reaction of school personnel to over-testing
   6. New tests and testing techniques
   7. Effects of coaching

4. Business of Testing
   1. Names of test publishers and national testing programs
   2. Research and advisory activities
   3. Scoring services
   4. The overlap of testing programs

5. Procedures in Making a Standardized Test
   1. Use of advisory specialists
   2. Experimental testing (pre-testing and research)
   3. Norming procedures

6. Programmed Learning Using Testing Procedures
   1. Defined
   2. Linear
   3. Branching
   4. Teaching machines

Appendix B

Items in Practice, Unit, and Final Examination Classified According To Subjects Covered

The two digits in the left hand column for each test indicate major subject matter classification and sub-classification as given in Appendix A. The next two columns under Practice Test and Unit Test indicate the second sub-classification for each item. The last column indicates the number of items in the Unit where there is no item coded the same as in the Practice Test. For example, in the second row under the column headed "Code Classification" the number "02" is found. Referring to **Appendix A**, the "0" represents the major code classification of "Scoring Systems" and the "2" represents the first sub-classification "Central Tendencies." In the row reading across from the "02", the next column under Practice Test is blank. A "2" is found under "Unit Test." This "2" represents a further sub-classification, "Median." A "1" is listed in the last column to indicate that under the major classification "02" there is one item in the Unit Test lacking a corresponding parallel item in the Practice Test. The remainder of the table may be interpreted in a similar manner. It should be kept in mind that each of the several numbers found in the Practice Test and the Unit Test columns is a further sub-classification of its corresponding test item.

TEST 1

Subject Matter Classification	Practice Test	Unit Test	Number of Items in Unit Test with No Parallel Coded Item in Practice Test
01	2		
02		2	1
03	1223333	0236	2
04	112	02	1
05			
06		0	1
07	000	0000	
08			
09	23		
10			
11	2	33	2
12			
13	1111133333	1112333333	1
14	11122	12222	
15	01122345	0111122445	
			Total 8

TEST 2

Subject Matter Classification	Practice Test	Unit Test	Number of Items in Unit Test with No Parallel Coded Item In Practice Test
21	12234	12222344	
23	00333455	023335	
24	00		
25	01122335668888	111225688888	
	999	999	
26	1244	12445	1
27			
28	03	267777	6
29			
30			
31			
32			
			Total 7

TEST 3

31	011	00	
32		1	1
33	2	0	1
34	0	11	2
35	0455	45	
36	011224	00022224	
37	1222222222	222222222	
38	2	3	1
39			
40			
41	11333	0233	2
42	0223	1223	1
43	01	01123	2
44	224	4	
			Total 10

TEST 4

51	00044	011	2
52	122	222	
53	23	223	
54	0122244	112224	
55	00333	0233	1
56		0	1
57	122	1222	
58			
59			
60			
61		3	1
62	23	3	
63	22444466	34444667	2
64	01	0345	3
65	0024	0	
66			
67		2	1
			Total 11

TEST 5

Subject Matter Classification	Practice Test	Unit Test	Number of Items in Unit Test with No Parallel Coded Item In Practice Test
71	23	4	1
72	0	3	1
73			
74	133	01	1
75	06	06	
76	000	0000	
77			
78			
79			
80			
81			
82			
83	01	01	
84	02233	22223	
85	1222	1244	2
86	1		
87		00	2
88			
89			
90			
91	000111112	01112	
92	00	011	2
93	6	267	2
94		0	1
95	2	2	
96	1234	2344	
			Total 12

Appendix C

Items in Practice, Unit and Final Examination Classified According To Skill Required For Response

Basic Skill Classification

	Practice Tests	Unit Tests	Final Examination
Factual knowledge	74	69	26
Application of knowledge	111	110	47
Evaluation of practice	15	21	7

Secondary Skill Classification*

	Practice	Unit Tests	Final Examination
Verbal reasoning	79	80	29
Statistics or computational	14	14	9
Both verbal and computational	15	28	11

*Many of the items, such as those requiring "factual knowledge" were not given a secondary skill classification.

Appendix D

Practice, Unit, and Final Examinations According To Difficulty and Biserial r

STATISTICAL CHARACTERISTICS OF THE TESTS*

Tests	Distribution of items according to difficulty level (Percent Correct)								Distribution of items according to Biserial r								
	Difficult			Average		Easy											
	20	30	40	50	60	70	80	90	.10	.20	.30	.40	.50	.60	.70	.80	.90
Practice Test 1	2	3	6	7	13	6	1	....	....	....	2	5	6	6	9	11	1
Unit Test 1	2	5	5	9	12	3	3	1	1	1	4	6	2	9	6	10	1
Practice Test 2	3	7	7	9	4	3	3	3	3	1	12	3	9	6	2	4	....
Unit Test 2	4	....	13	8	6	5	2	2	....	....	9	5	9	11	4	2	....
Practice Test 3	....	....	7	2	8	11	4	8	....	3	8	8	8	9	2	2	....
Unit Test 3	2	1	5	3	6	9	3	10	....	4	8	10	2	9	5	2	....
Practice Test 4	1	5	7	2	7	6	8	4	1	3	6	9	5	11	3	2	....
Unit Test 4	2	7	4	5	5	5	7	5	....	2	15	4	3	9	4	3	....
Practice Test 5	4	3	6	4	10	7	4	1	....	....	8	8	13	3	5	2	1
Unit Test 5	2	6	6	4	11	4	3	4	....	3	9	6	4	8	5	3	2
Final Examination																	
Items 1-45	2	1	6	4	10	7	10	4	3	6	3	7	9	10	3	2	1
Items 46-80	3	4	5	4	5	9	4	1	4	11	4	6	8	2	....	....	....

*Based on pretest results

Appendix E

Individual Tutoring Given to Reference and Experimental Groups By Sex

	Men	Women	Total
Reference			
Number in group	30	34	64
Number receiving tutoring	6	11	17
Percent of group	20.00	32.3	26.4
Number of sessions	12	40	52
Mean number of sessions per group	.40	1.18	.81
Experimental			
Number in group	44	37	81
Numbering receiving tutoring	13	7	20
Percent of group	29.5	18.9	24.7
Number of sessions	35	17	52
Mean number of sessions per group	.80	.46	.64

Appendix F

Differences Between Practice and Unit Tests

Unit Number	Practice Test Mean	Unit Test Mean	Practice Test SD	Unit Test SD	Correlation Coefficient between two Tests	"t" of Difference
I	17.17	22.54	6.16	6.39	.623	8.57***
II	18.88	21.04	4.85	5.30	.696	4.73***
III	26.89	28.24	4.75	3.91	.510	2.69***
IV	22.82	26.20	5.45	4.71	.671	7.22***
V	19.61	20.13	6.01	4.78	.670	.99

***=Significant at .01 level

Appendix G

Intercorrelation Among Test Scores

Intercorrelation Among Test Scores For Total Group

N=145	1	2	3	4	5	Finals	Sums 1-5	Total	SCAT
Unit Test 1		.58	.58	.62	.56	.70	.85	.84	.52
Unit Test 2			.55	.53	.50	.59	.80	.77	.45
Unit Test 3				.53	.46	.64	.76	.76	.56
Unit Test 4					.57	.65	.80	.79	.43
Unit Test 5						.60	.75	.74	.37
Finals							.79	.90	.49
Sums 1-5								.98	.58
Total (1-5 and Final)									.58

Intercorrelation Among Test Scores For Reference Group

N=145	1	2	3	4	5	Finals	Sums 1-5	Total	SCAT
Unit Test 1		.50	.61	.69	.60	.57	.89	.84	.47
Unit Test 2			.40	.40	.37	.41	.71	.66	.36
Unit Test 3				.50	.41	.56	.74	.73	.47
Unit Test 4					.49	.60	.80	.79	.37
Unit Test 5						.54	.73	.72	.42
Finals							.69	.84	.30
Sums 1-5								.97	.53
Total (1-5 and Final)									.50

APPENDIX G—(Continued)

Intercorrelation Among Test Scores For Experimental Group

N=145	1	2	3	4	5	Finals	Sums 1-5	Total	SCAT
Unit Test 1		.61	.55	.57	.54	.75	.83	.84	.52
Unit Test 2			.63	.60	.59	.68	.85	.82	.48
Unit Test 3				.52	.50	.68	.77	.77	.60
Unit Test 4					.62	.67	.79	.78	.45
Unit Test 5						.62	.77	.75	.33
Finals							.83	.92	.59
Sums 1-5								.98	.59
Total (1-5 and Final)									.61

Appendix H

Summary of Analyses of Variance and Covariance Results for Reference and Experimental Groups Separated into Men and Women

Reference Group

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Men	Women	F	Men	Women	Men	Women	F	Men	Women	F
Test 1	19.90	19.65	.04	19.61	31.14	.55	.48	.00	19.72	19.81	.01
Test 2	18.20	18.62	.11	19.68	32.06	.37	.45	.34	18.05	18.75	.35
Test 3	25.13	26.62	1.75	19.29	20.67	.72	.48	1.77	24.94	26.78	4.01**
Test 4	23.77	24.62	.49	17.22	28.85	.55	.32	.45	23.63	24.74	.96
Test 5	18.40	19.32	.74	20.94	16.23	.36	.57	.24	18.26	19.45	1.53
Final	40.37	43.21	1.74	79.00	69.56	.54	.32	1.47	40.11	43.43	2.82*
Sum of Tests 1-5	105.40	108.82	.53	289.77	406.15	.66	.63	.02	104.01	110.05	2.73
Sum of All Tests	145.77	152.03	.95	613.70	696.39	.65	.53	.34	144.71	152.96	2.43
SCAT	78.40	76.32	.29	215.70	248.99						

NUMBER OF STUDENTS:
Men—30
Women—34
Total—64

*—.10 Confidence Level
**—.05 Confidence Level

Appendix H—(Continued)

Experimental Group

	Unadjusted Means			Variance		Correlation with SCAT		Parallel Regression Lines	Adjusted Means		
	Men	Women	F	Men	Women	Men	Women	F	Men	Women	F
Test 1	22.64	21.68	.44	34.52	50.73	.20	.78	10.06***	22.60	21.71	.52
Test 2	20.95	20.86	.00	27.13	29.73	.32	.65	1.67	20.91	20.89	.00
Test 3	28.05	28.30	.08	11.30	20.27	.33	.80	8.71***	28.02	28.32	.19
Test 4	25.89	26.54	.39	18.10	27.31	.24	.61	3.36	25.87	26.57	.55
Test 5	19.30	20.70	1.79	16.49	29.21	.02	.61	3.00	19.32	20.67	1.86
Final	48.86	45.62	.33	63.52	129.63	.24	.82	14.45***	46.81	45.69	.41
Sum of Tests 1-5	116.82	117.81	.05	275.50	574.94	.32	.77	8.82***	116.67	117.99	.13
Sum of All Tests	163.68	163.43	.00	540.73	1187.31	.32	.81	11.81***	163.51	163.64	.00
SCAT	83.77	83.49	.01	141.99	216.85						

NUMBER OF STUDENTS:
Men—44
Women—37
Total—81

***—Significant at .01 level

APPENDIX I

Samples of Items in Which Either Experimental or Reference Group Scored Significantly Higher Than the Other at the .05 Level

For purposes of information and general interest, two items were selected on each of the five unit tests. On one item the experimental group scored higher and on the other the reference group scored higher. Generally speaking, these items were selected because they represented the greatest difference between experimental and reference in percent of correct responses. These differences were all significant at the .05 level.

In addition to listing these items from the unit tests, an attempt was made to pick out items from the practice tests which were reasonably parallel to the unit test items. As may be seen, in some cases the items in the practice test are very similar to those in the unit test. It is left to the reader's judgment to decide what advantage there might be to students who had an opportunity to attempt the practice item. In some cases no item was found in the practice test which could be designated as parallel to the one selected in the unit test.

With each item a partial code of item characteristics and item statistics is given.

Sample Items from Unit Test 1

Item on which Experimental Group Scored Significantly Higher Than Reference Group

A test with a low coefficient of predictive validity is ordinarily useful in a selection procedure if

A. almost all applicants are to be chosen
B. applicants with the best qualifications are to be chosen
C. applicants of lower ability are to be considered
D. very few applicants from a large number are to be chosen
E. very few applicants from a small number are to be chosen

Key D

CODE OF ITEM CHARACTERISTICS
Content:
1. Regression and correlation
5. Problems of interpreting and use of correlation coefficients and regression equations
5. Satisfactory correlation for selection procedure
Skill:
1. Factual knowledge

ITEM STATISTICS
Percent correct
reference—50%
experimental—90%
t (for difference) = 5.33
Biserial r
reference—.35
experimental—.21

Item in Practice Test Most Nearly Parallel to Unit Test Item

In selection of personnel, which of the following correlations between scores on predictor test and job success would be best for an employer in selecting job applicants.

A. .80, 70 applicants for 70 openings
B. .60, 60 applicants for 60 openings
C. .30, 60 applicants for 30 openings
D. .25, 25 applicants for 30 openings
E. .20, 1000 applicants for 20 openings

Key E

CODE OF ITEM CHARACTERISTICS
Content:
1. Regression and correlation
5. Problems of interpreting and use of correlation coefficient and regression equations
5. Satisfactory correlation for selection procedure
Skill:
2. Application of knowledge

ITEM STATISTICS
Percent correct
experimental—43%
Biserial r
experimental—.55

Item on which Reference Group Scored Significantly Higher Than Experimental Group

The mean score of a normal distribution of test scores is 50. The standard deviation is 8. Which of the following would increase both the mean and standard deviation of a set of test scores?

A. Adding 2 to each test score
B. Multiplying each test score by 2

2 1

C. Subtracting 2 from each test score
D. Dividing each test score by 2
E. Adding 50 to each test score

Key B

CODE OF ITEM CHARACTERISTICS
Content:
0. Scoring systems
4. Dispersion, spread or position (based on concept of normal distribution)
3. Standard deviation (including effects of adding, subtracting, etc.)
Skill:
2. Application of knowledge
3. Both verbal and computational
Special characteristics:
5. Extraneous data

ITEM STATISTICS
Percent correct
reference—81%
experimental—66%
t (for difference) = 2.00
Biserial r
reference—.72
experimental—.74

Item in Practice Test Most Nearly Parallel to Unit Test Item

Which of the following will probably increase the standard deviation of a group of test scores the most?

A. Add 1000 to each score
B. Multiply each score by .59
C. Add .1001 to each score
D. Multiply each score by 1.001
E. Divide each score by .59

Key E

CODE OF ITEM CHARACTERISTICS
Content:
0. Scoring systems
4. Dispersion, spread or position (based on concept of normal distribution)
3. Standard deviation (including effects of adding, subtracting, etc.)
Skill:
2. Application of knowledge
2. Statistical or computational

ITEM STATISTICS
Percent correct
experimental—.23
Biserial r
experimental—.33

Items on which Experimental Group Scored Significantly Higher than Reference Group

The following data is the basis for the next questions.

	Column One Observed score		Column Two Estimated true score
(A)	+40	.	+40
	+30	.	+30
(B)	+20	.	+20
	+10	.	+10
(C)	0	.	0
(D)	−10	.	−10
	−20	.	−20
	−30	.	−30
(E)	−40	.	−40

In order to answer the questions, draw the following lines from column one to column two.

A. +30 to +40 B. 0 to 0 C. −10 to 0
D. −30 to −30 E. −40 to −30

Which line most certainly represents a reliability of .00?

Key C

CODE OF ITEM CHARACTERISTICS

Content:

2. Evaluation of a test—validity and reliability
   5. The concept of reliability
      7. Reliability coefficient as regression line (true score-observed score relationship)

Skill:

2. Application of knowledge
   3. Both verbal and computational

Special Characteristics:

2. Use of chart or table

ITEM STATISTICS

Percent correct
reference—38%
experimental—85%
t (for difference) = 5.87

Biserial r
reference—.55
experimental—.28

Item in Practice Test Most Nearly Parallel to Unit Test Item

The following data is the basis for the next questions.

	Column One Observed score		Column Two Estimated true score
(A)	+40	.	+40
	+30	.	+30
(B)	+20	.	+20
	+10	.	+10
(C)	0	.	0
(D)	−10	.	−10
	−20	.	−20
	−30	.	−30
(E)	−40	.	−40

In order to answer the questions, draw the following lines from column one to column two.

A. +30 to +40 B. +20 to 0 C. +10 to +10
D. 0 to 0 E. −40 to −10

Which line most certainly represents a reliability of .00?

Key B

CODE OF ITEM CHARACTERISTICS

Content:

2. Evaluating a test—validity and reliability
   5. The concept of reliability
      7. Reliability coefficient as regression line (true score-observed score relationship)

Skill:

2. Application of knowledge
   3. Both verbal and computational

Special characteristics:

2. Use of chart or table

ITEM STATISTICS

Percent correct experimental—93%

Biserial r—.25

Item on which Reference Group Scored Significantly Higher Than Experimental Group

The following students are listed in order of their scores on a test with the best student first. The letters (A B C D E) represent test questions. R means the student got the question right. W means the student got the question wrong. Which question, if eliminated, will raise the reliability of the test the most using the inter-item relationship method of computing reliability?

	A	B	C	D	E
John	R	W	R	R	R
Mary	R	W	R	W	R
Susy	W	R	W	W	R
Bill	W	R	W	W	R
Alice	R	R	W	W	W

Key B

CODE OF ITEM CHARACTERISTICS

Content:

2. Evaluating a test
   5. The concept of reliability
      5. Item intercorrelations (Kuder-Richardson)

Skill:

2. Application of knowledge
   3. Both verbal and computational

Special Characteristics:

2. Use of chart or table

ITEM STATISTICS

Percent correct

reference—59%
experimental—32%
t (for difference) = 3.29

Biserial r

reference—.24
experimental—.33

No item was identified in practice test as being even somewhat parallel.

Item on which Experimental Group Scored Significantly Higher Than Reference Group

The following item analysis table is for the next questions.

	A	B*	C	D	E
H	19	25	40	15	1
L	24	3	20	50	3

Which distractor is probably least in need of revision?

Key D

CODE OF ITEM CHARACTERISTICS

Content:

3. Writing a classroom test
   7. Evaluating test questions
      2. Item analysis

Skill:

2. Application of knowledge
   3. Both verbal and computational

Special Characteristics:

2. Use of chart or table

ITEM STATISTICS

Percent correct

reference—28%
experimental—74%
t (for difference) = 6.66

Biserial r

reference—.43
experimental—.71

Item in Practice Test Most Nearly Parallel to Unit Test Item

The next questions are based on the following item analysis.

Item	Omit	1	2	3	4	5	NR	TT
A. High	0	90*	5	2	2	1	0	100
A. Low	0	25	50	10	15	0	0	100
B. High	8	20	12	11	19	30*	0	100
B. Low	9	16	18	13	15	29	0	100
C. High	0	92*	8	0	0	0	0	100
C. Low	10	10	15	15	20	30	0	100
D. High	0	5	5	75*	5	10	0	100
D. Low	9	20	20	25	10	15	1	99
E. High	15	5	17	8	35*	18	2	98
E. Low	0	5	20	10	10	15	40	60

0 = Omit
1 - 5 = Option numbers
NR = Number not reaching item
TT = Total number trying this item
* = Correct answer

Which item is most likely to be ambiguous?

Key B

CODE OF ITEM CHARACTERISTICS

Content:

3. Writing a classroom test

7. Evaluating test questions

2. Item analysis

Skill:

2. Application of knowledge

3. Both verbal and computational

Special characteristics:

2. Use of chart or table

ITEM STATISTICS

Percent correct
experimental—.57

Biserial r
experimental—.46

Item on which Reference Group Scored Significantly Higher Than Experimental Group

Which one of the adverse criticisms best describes the test question below designed to test computational and applied problem solving ability:

> "If an individual who is engaged in the retail distribution of automobiles carries through to completion a business transaction wherein he purchases a motor vehicle for the sum of two thousand five hundred twenty-five dollars and at a subsequent date disposes of the said vehicle for the sum of one thousand six hundred eighty-three dollars, what is the net profit or loss on the completed business transaction involving the given vehicle?"

A. The question is ambiguous in meaning.
B. The question has a specific determiner.
C. The question tests an unimportant detail.
D. The question presents a reading difficulty.
E. The question does not follow accepted English usage.

Key D

CODE OF ITEM CHARACTERISTICS

Content:
3. Writing a classroom test
6. Writing the test item
2. State questions clearly

Skill:
2. Application of knowledge
1. Verbal

ITEM STATISTICS

Percent correct
reference—94%
experimental—74%
t (for difference) = 3.39

Biserial r
reference—.18
experimental—.34

Item Most Nearly Parallel in Practice Test

The stem of this question is (referring to the question you are now reading)

A. weak because it does not specify the task
B. an illustration of grammatical parallelism
C. undesirable because it is ambiguous
D. a good stem for a multiple-choice item
E. an example of stereotype wording

Key A

CODE OF ITEM CHARACTERISTICS

Content:
3. Writing a classroom test
6. Writing the test item
2. State question clearly

Skill:
2. Application of knowledge
1. Verbal

ITEM STATISTICS

Percent correct
experimental—43%

Biserial r
experimental—00

Sample Items from Unit Test 4

Item on which Experimental Group Scored Significantly Higher Than Reference Group

If one of the following was to be used as a predictor test for a class of students in school which one would be the most satisfactory?

A. A non-verbal test
B. A test based on factor analysis
C. A diagnostic test
D. A test of verbal abilities
E. An individual type mental abilities test

Key D

CODE OF ITEM CHARACTERISTICS

Content:
5. Standardized "ability" tests
3. Group tests
2. Advantages

Skill:
3. Evaluating of practices
1. Verbal reasoning

ITEM STATISTICS

Percent correct
reference—25%
experimental—63%
t (for difference) = 7.60

Biserial r
reference—.02
experimental—.58

Item in Practice Test Most Nearly Parallel to Unit Test Item

Ordinarily, which one of the following is the most accurate and practical for predicting school success?

A. An individual mental abilities test
B. A test of verbal abilities
C. A test of mathematical abilities
D. A non-verbal test
E. A performance test

Key B

CODE OF ITEM CHARACTERISTICS

Content:
5. Standardized "ability" tests
5. Breakdown of general to separate abilities
1. Verbal reasoning

Skill:
3. Evaluating of practices
1. Verbal reasoning

ITEM STATISTICS

Percent correct
experimental—26%

Biserial r
experimental—.31

Item on which Reference Group Scored Significantly Higher Than Experimental Group

How would you explain the following statement: "The average percentile score for a class is 60. This class is at the 99th percentile?"

A. The 60 must refer to a raw score.
B. An average score of 60 is very high.
C. Only the best students in the class are at the 99th percentile.
D. An average score of 60 is really low.
E. A mistake in computation has been made.

Key B

CODE OF ITEM CHARACTERISTICS

Content:
6. Standardized "achievement" tests
3. Interpretation of scores
7. Group norms compared with individual norms

Skill:
2. Application of knowledge
3. Both verbal and computational

ITEM STATISTICS

Percent correct
reference—83%
experimental—52%
t (for difference) = 4.70

Biserial r
reference—.36
experimental—.12

No item was identified in the practice test as even being somewhat parallel.

Sample Items from Unit Test 5

Item on which Experimental Group Scored Significantly Higher Than Reference Group

The following indicate scales on the MMPI and are to be used as options for the next questions.

A. The question score
B. The lie score (L)
C. The validity score (F)
D. The correction score (K)
E. None of the above

If a person was reasonably honest but tended to present himself in a generally favorable light, on what scale would he get a high score.

Key D

CODE OF ITEM CHARACTERISTICS

Content:
8. Measuring personality
4. The MMPI (a personality inventory)
2. "Validity" scores

Skills:
2. Application of knowledge
1. Verbal reasoning

ITEM STATISTICS

Percent correct
reference—7%
experimental—48%
t (for difference) = 6.61

Biserial r
reference—00
experimental—.53

Item in Practice Test Most Nearly Parallel to Unit Test Item

The following options are for the next question.

A. L scale
B. K scale
C. MF scale
D. Questions mark scale
E. F scale

A person who was trying to look bad on the MMPI would tend to get a high score on which scale?

Key E

CODE OF ITEM CHARACTERISTICS
Content:
8. Measuring personality
4. The MMPI (a personality inventory)
2. "Validity" scores
Skills:
2. Application of knowledge
1. Verbal reasoning

ITEM STATISTICS
Percent correct
experimental—24%
Biserial r
experimental—.37

Item on Which Reference Group Scored Significantly Higher Than Experimental Group

The following options are for the next question.

A. Skinner
B. Crowder
C. Pavlov
D. Hull
E. Pressey

Who of the preceding is most closely associated with linear programming?

Key A

CODE OF ITEM CHARACTERISTICS
Content:
9. History of and current practices in testing
6. Programmed learning using testing procedures
2. Linear
Skills:
1. Factual knowledge

ITEM STATISTICS
Percent correct
reference—57%
experimental—20%
t (for difference) = 5.69
Biserial r
reference—.40
experimental—.45

Item Most Nearly Parallel in Practice Test

The following options are for the next question.

A. Skinner B. Crowder C. Pavlov
D. Hull E. Pressey

Who of the preceding is most closely associated with branching?

Key B

CODE OF ITEM CHARACTERISTICS

Content:

9. History of and current practices in testing
6. Programmed learning using testing procedures
3. Branching

Skills:

1. Factual knowledge

ITEM STATISTICS

Percent correct
experimental—55%

Biserial r
experimental—.53